Assessment Schedule – 2009

Scholarship Statistics and Modelling (93201)

Evidence Statement

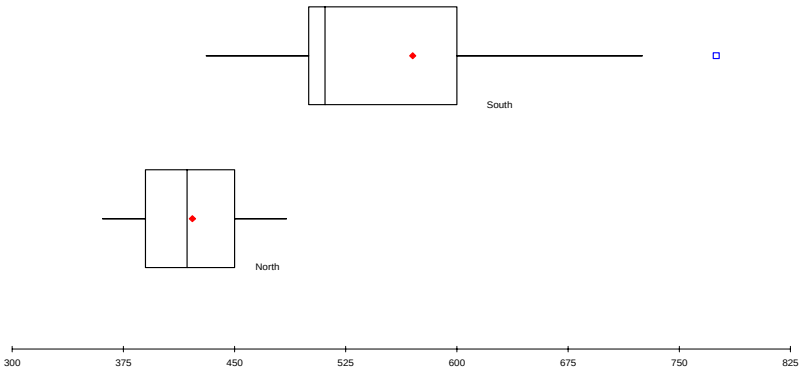
General Principles:

1. Ignore incorrect answers if alongside correct answers. The exception is contradictory statements.
2. Ignore minor copying errors.

QUESTION ONE

Tasks Q1 (a)

Evidence:
Box-Plot Showing Price Comparisons in ($000) for North and South Suburbs



Distribution of Valuations:
North – roughly symmetrical with skew = $3 000
South – Skewed to the right (positive skew, higher values) with skew = $102 000

Ranges:
North – ($361 000 to $485 000) i.e. $124 000
South – ($431 000 to $775 000) i.e. $344 000 with an outlier at $775 000

Inter-quartile Ranges:
North – ($390 000 to $450 000) i.e. $60 000
South – ($500 000 to $600 000) i.e. $100 000

Central Measures:
Median - $93 000 higher for South
Mean - $192 000 higher for South

Spread:
The standard deviation is $15 000 higher for South so valuations are more spread out about the mean.

Overlap of Boxes:
The top 75% of home valuations from the South Suburb have a higher valuation than any home from the North Suburb.

Note:

1. All statements have to be comparative and distinct.
2. Can have other statements.
3. A general statement must be backed up by descriptive statistics.
4. Can ignore direction of skew.
5. Graph must distinguish between North and South Suburbs and have a scale.
6. Must be box-plots.
7. Approximate positioning of key points acceptable.
8. Do not need to show mean or outlier.

Tasks Q1 (b)

Evidence:
Use stratified random sampling with proportional allocation. Select 80 homes at random from the North suburb and 120 homes at random from the South suburb. Allocate each home a five digit random number according to the listing on the database. Use five digit numbers to make selections from each suburb ignoring repetitions and numbers outside range.

Note:
1. Random sampling on 80 000 not acceptable.
2. First two sentences in evidence not complete enough.
3. Systematic sampling within each suburb with k = 400 (must show method for finding the value of k) is acceptable.

Tasks Q1 (c)

Evidence:
The 95% confidence interval for the difference in the population mean valuations is given by:

$$\mu_1 - \mu_2 = (613 - 421) \pm 1.96\sqrt{\frac{50^2}{120} + \frac{35^2}{80}}$$

So \$180 200 < $\mu_1 - \mu_2$ < \$203 800 or $\mu_1 - \mu_2 = \$192{,}000 \pm \$11{,}800$

Note:
1. Rounding errors accepted.
2. Can have $\mu_1 - \mu_2$ = -\$192 000 $\pm$ \$11,800

Tasks Q1 (d)

Evidence:
As zero is not contained within the confidence interval we conclude that there is a difference in the mean valuation between the two suburbs, North and South.

Note:
1. Ignore use of word "significance" and ignore no \$ sign.
2. Evidence must be in context.
3. Can have, mean valuations in the South suburb are higher than the mean valuations in the North suburb.

Tasks Q1 (e)

Evidence:

Overall Mean Estimate = $\dfrac{421 \times 80 + 613 \times 120}{200}$ = \$536 200

Note:
1. No \$ MEI
2. Some recognition of \$000 required in either (c), (d) or (e). Only penalised once.

Tasks Q1 (f)

Evidence:

$$n = \left(\frac{z\sigma}{E}\right)^2 = \left(\frac{2.576 \times 97000}{9000}\right)^2 = 771$$

So number of further valuations that should be sampled from:
North is (32/80) x 771 – 80 = 228 and
South is (48/80) x 771 – 120 = 343

Note:
Accept minor differences i.e. 230 from North and 345 from South (multiples of five).

Judgement:

(a) **S**: Correct graph plus three comparative statements. **P**: Graph only.
(b), (c) & (d) **S**: Any two correct. **P**: Any one correct.
(e) & (f) **O**: Both correct. **P**: Any one correct.

Tasks Q2 (a)

Evidence:

(i) Choose model based on best fit to a line on a transformed graph: ln *S* versus ln *D* or ln *S* versus *D* The best model is the exponential model with $S = 601.7e^{-0.0433D}$ It's the best fit according to the plot and is shown by the points lining up with a fitted line having virtually no residuals.

(ii) Predicted selling price is $601.7e^{-0.0433x37}$ = \$121 000. Despite the fact that the exponential model fits very well, D = 37 is well outside the range of the given data which only consists of eight points so prediction is suspect.

(iii) Main limitation is that the non-linear model does not take into account other factors that would influence the selling price S like location of home, size of home, age of home,…..

Note:

1. Not correct if values of constants generated from **ANY** two points.
2. Prediction can be correct if consistent with b and k values.
3. Must have two or more factors to be correct in (iii).
4. Variation in the values of constants acceptable.
5. Nature of exponential model explained is acceptable in either (ii) or (iii) but not in both.
6. Part (ii) can be correct with carried error.
7. R^2 values okay for justification of best model.

Judgement

2S: Any two parts correct.
S + P: Any one part correct and at least one other part partially correct.
S: Any one part correct.
2P: Any two parts partially correct.
P: Any one part partially correct.

Note:

1. Partially correct in (i) implies that only justification or constants are correct.
2. Partially correct in (ii) implies that the validity point is correct or only the predicted price is correct.
3. Can have a carried error for the prediction from (i).
4. Partially correct in (iii) implies that only one factor is correct.

Tasks Q2 (b)

Evidence:
Let P = event home has a pool
Let B = event home has a BBQ

Set x = pr (P only), y = pr (P $\cap$ B) and z = pr (B only)

Now x + y + z = 0.95, $\frac{y}{x+y} = 0.8$ and $\frac{y}{y+z} = 0.4$
We need to solve these simultaneous equations for y:
$x = \frac{1}{4}y$ and $z = \frac{3}{2}y$ so $\frac{1}{4}y + y + \frac{3}{2}y = 0.95$ which gives $y = 0.3455$
So % of homes that have both a pool and an outdoor BBQ area is 34.55%.

Judgement

O: Correct answer.
P: Appreciation shown of the use of conditional probability.

Note:

1. No \$ sign MEI
2. Accept 34.5 or 34.6.
3. Answer of 0.3455 acceptable as MEI

Tasks Q3 (a)

Evidence:

A linear positive relationship is indicated for South suburb (ex outlier). There is an outlier at ($190m^2$, \$820 000) for South suburb. There is little scatter for South Suburb indicating a strong relationship. A non – linear relationship seems more likely as a fit for the North suburb which has two distinct points with the least floor area separate from the other points. Three gradients are evident in the North suburb scatter; slightly positive, very steep and positive.

Note:

1. Must specify outlier.
2. Can have other reasonable points i.e. interpretation of gradient of fitted line.
3. Accept answers with 000 omitted.

Judgement

S: Three correct points about the relationship with both graphs being commented on.
P: One or two correct points about the relationship.

Tasks Q3 (b)

Evidence:

Equation 2 is more appropriate for North due to the non-linear position of points. Equation 3 is more appropriate for South due to closeness of fit of line to points compared to equation 4.
For A = 255, predicted selling price for North = $0.071(255)^2 – 31.214(255) + 3813$ = \$470 000
For A = 255, predicted selling price for South = $4.9163(255) – 505.9$ = \$748 000

For South, validity is good as variability near A = 255 is low and correlation is reasonably high. However for North, A = 255 is in a gap where there is a large variance in the selling price so validity is suspect. A reservation with both predictions is that the models are based only on 16 points from each suburb.

Note:

1. Must have \$000 in predictions at least once otherwise penalty of one answer count.
2. A validity comment about each prediction is required. For south, must have both good correlation **AND** close to data.
3. Cannot repeat comments from (a).
4. If equation wrong, prediction doesn't count.
5. MEI – Wrong choice of equation but correct one substituted into for prediction.

Judgement

O: Five or six answers correct out of justify, calculate prediction and validity comment (3 x 2 = 6 answers).
P: Three or four answers correct out of the above.

Tasks Q3(c)

Evidence:

For "**age of home**", selling price would be influenced by such factors as type of home, the era in which the home was built, current condition of the home and the style of home. For "**location of home**" the selling price would be influenced by factors such as demand for homes in its particular location. This demand would be influenced by factors such as rural or urban preferences for location, availability of employment, reputation of area, desirable lifestyle options and perceived resale value. If the factors are favourable, in each case, then the selling price of the home would be expected to be higher.

Judgement

S: Four points or more correct.
P: Two or three points correct.

Note:

1. Require at least two points about age and two points about location.
2. Two points, one positive and one negative about location can count as two points under location.

Task Q4 (a)

Evidence:

1. Median home prices have increased steadily until the September 07 quarter then started to drop then for the last two quarters June and September 09, prices have started to increase again.
2. Prices seem lower in the first quarter of each year due to seasonal variations.
3. The largest drop in prices occurred between the December 08 and March 09 quarters.
4. There was an average increase of $4 600 per quarter between March 04 and September 07.
5. There was an average decrease of $5 700 per quarter between December 07 and March 09.
6. The median home prices peaked at $351 000 in the September 07 quarter.

Judgement

S: Three or more points correct.
P: One or two points correct**.**

Task Q4 (b)

Evidence:
Choose equation 3 as it reflects the most recent trend (R^2 value close to 1)
For March 2010, $x = 25$
Trend Value = -5.66(25) + 437 = 295.5
For March quarter, S = $\frac{-13+-8+-8-2-16}{5}$ = –9.4 or S = $\frac{-2-16}{2}$ = -9.0
Forecast = 295.5 – 9.4 = 286.1 ie $286 100
Prediction is probably invalid as trend line used doesn't reflect the most immediate upward trend in the June and September 09 quarters.

Judgement

S: Justification of method**,** forecast and validity comment correct.
P: Any one of choice of equation justification, forecast or validity comment correct. Must choose equation 3 otherwise can only score on the validity point.
N: Not acceptable if last three points are used to fit line.

Task Q4(c)

Evidence:

June 2004: Index Number = $\frac{243}{284} \times 1000 = 856$ compare to Inflation Index = 924

June 2005: Index Number = $\frac{284}{284} \times 1000 = 1000$ compare to Inflation Index = 1000

June 2006: Index Number = $\frac{310}{284} \times 1000 = 1092$ compare to Inflation Index = 1047

June 2007: Index Number = $\frac{347}{284} \times 1000 = 1222$ compare to Inflation Index = 1208

June 2008: Index Number = $\frac{340}{284} \times 1000 = 1197$ compare to Inflation Index = 1224

June 2009: Index Number = $\frac{314}{284} \times 1000 = 1106$ compare to Inflation Index = 1249

Between June 2004 and June 2007, the median home prices have increased at a faster rate than inflation. That is "increased in real terms". In June 2008 and June 2009, the median home prices have fallen in "real terms" (slower rate than inflation) compared to June 2005

Judgement

O: Correct Index numbers and both comparisons (increase/ decrease) correctly made of the two series.
P: Correct Index numbers or correct explanation based on "deflated" home price values

Note:

1. Comments like faster than inflation or slower than inflation acceptable.
2. MEI for slight calculation errors.

Tasks Q5 (a) (i), (a) (ii) and (b)

Evidence (a) (i):

$n = 50$, $p = 0.3$ so $\mu = np = 15$ and $\sigma = \sqrt{np(1-p)} = \sqrt{(50 \times 0.3 \times (1-0.3))} = 3.24$

So pr($x \leq 20$) continuity correction = pr ($x < 20.5$) = pr ($z \leq 1.697$) = 0.5 + 0.4552 = **0.9552**

Note:
Exact answer using binomial distribution directly with graphic calculator is **0.9522**.

Judgement

S: Correct answer.
P: Recognition of the use of the Binomial distribution or answer 0.9386 where continuity correction hasn't been applied.

Evidence (a) (ii):

n = 50, $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{34}{\sqrt{50}} = 4.808$

So pr ($\bar{x} > 250$) = pr ($z > 1.0398$) = 0.5 – 0.3508 = **0.1492**

Note:

1. Can use totals where mean = 12,250, std deviation = $34\sqrt{50}$ and pr (total > 12,500) is required.

Judgement

O: Correct answer.
P: Recognition of the use of the sampling distribution of the mean.

Evidence (b):
Method 1: Use Binomial directly

Get $(0.98)^n = 0.6$ so $n = \frac{\log 0.6}{\log 0.98} = 25$

Method 2: Use Poisson formulae

$e^{-\lambda} = 0.6$ so $\lambda = 0.52$ Hence $n = \frac{0.52}{0.02} = 26$

Method 3: Poisson approx to the Binomial

$\lambda = np = 0.02n$. When pr(x=0) = 0.6 then $\lambda = 0.5$ so $n = \frac{0.5}{0.02} = 25$

Judgement

S: Correct answer using any method (accept 25 or 26).
P: Selection of the appropriate probability distribution(s).

Tasks Q6 (a)

Evidence:
Let x = number of single detached homes
Let y = number of duplex units
The constraints are:

- $600x + 800y \leq 60\,000$ when simplified gives. $3x + 4y \leq 300$
- $65x + 100y \leq 7\,000$ when simplified gives $13x + 20y \leq 1400$

The profit function is $P = k(x + 1.5y)$ where k is a constant.

To ensure maximum profit ($P_{MAX} = 106k$ in each case):

$x = 40$ and $y = 44$ that is **40 single detached homes and 44 duplex units** should be built **or**
$x = 43$ and $y = 42$ that is **43 single detached homes and 42 duplex units** should be built **or**
$x = 46$ and $y = 40$ that is **46 single detached homes and 40 duplex units** should be built **or**
$x = 49$ and $y = 38$ that is **49 single detached homes and 38 duplex units** should be built **or**
$x = 52$ and $y = 36$ that is **52 single detached homes and 36 duplex units** should be built.

Judgement

S: Any one of these optimal point(s).
P: All two constraints correct OR any constraint correct plus the correct profit objective function.
P: Feasible region clearly shown on graph as opposed to constraint equations.

Note:

1. Can use graph as evidence for the constraints.
2. Profit function can take the form $P = 1.0x + 1.5y$.

Tasks Q6 (b)

Evidence:

New Council regulation: $3y \leq 2x$ or $y \leq \frac{2}{3}x$

Now to ensure maximum profit, $x = 53$ and $y = 35$ **or** $x = 56$ and $y = 33$.

For either optimal point, % reduction in maximum profit = $\frac{0.5k}{106k} \times 100$ = **0.47%.**

Judgement

S: Correct method and answer or correct new optimal point and carried error in % reduction in profit due to wrong answer in (a).
P: Identification of correct new optimal point
N: Incorrect new optimal point.

Note:

1. Only one optimal point is required.

Tasks Q6(c)

Evidence:
With social constraint, point where profit is maximised is given by: $x = 85$ and $y = 11$ with profit = \$101.5k

Unused land = $60\,000 - 800(11) - 600(85)$ = **200 m**^{2}
Unused budget = $7\,000 - 100(11) - 65(85)$ = **\$375 000**

Judgement

O: Correct method and answers for both land and budget.
P: Correct point for maximising profit or consistent answers on incorrect point.

Scoring for each Question

The codes are put together for each question and then converted to a mark out of eight according to the following table:

Mark	Codes
8	O + 2S
7	O + P, O + 2P, O + S, O + S + P
6	2S, 2S + P
5	S + P, S + 2P
4	S
3	3P
2	2P
1	P
0	N

The marks for each question are totalled to give an overall mark. Best possible overall mark is 48.